Published online 18 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W79–W85***
doi:10.1093/nar/gkr291

Genome surveyor 2.0: *cis*-regulatory analysis in *Drosophila*

Majid Kazemian[1], Michael H. Brodsky[2,4] and Saurabh Sinha[1,3,*]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, [2]Program in Gene Function and Expression, University of Massachusetts Medical School, Worcester, MA 01655, [3]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 and [4]Department of Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01655, USA

Received February 3, 2011; Revised April 7, 2011; Accepted April 13, 2011

ABSTRACT

Genome Surveyor 2.0 is a web-based tool for discovery and analysis of *cis*-regulatory elements in *Drosophila*, built on top of the GBrowse genome browser for convenient visualization. Genome Surveyor was developed as a tool for predicting transcription factor (TF) binding targets and *cis*-regulatory modules (CRMs/enhancers), based on motifs representing experimentally determined DNA binding specificities. Since its first publication, we have added substantial new functionality (e.g. phylogenetic averaging of motif scores from multiple species, and a novel CRM discovery technique), increased the number of supported motifs about 4-fold (from ~100 to ~400), added provisions for evolutionary comparison across many more *Drosophila* species (from 2 to 12), and improved the user-interface. The server is free and open to all users, and there is no login requirement. Address: http://veda.cs.uiuc.edu/gs.

INTRODUCTION

Cis-regulatory analysis is a key step in understanding and decoding transcriptional regulatory networks. The researcher is interested in determining which transcription factors (TFs) regulate a gene (or genes) of interest, the locations of binding sites for those TFs, and, if the analysis has an evolutionary component, how those binding sites and regulatory influences evolve across species. For *Drosophila* researchers, these tasks have been greatly facilitated by the availability of 12 *Drosophila* genomes (1,2) and vast amounts of other genetic and genomic data (3–5). In addition, a variety of computational tools can nicely complement high throughput experimental approaches to the above tasks, and aid the biologist to efficiently design and conduct hypothesis-driven experiments. For instance, available computational methods can summarize known binding specificities of TFs as 'motifs' and search the genome (or genomic regions near a specific gene) for matches to these motifs, thus identifying putative TF binding sites. Other more sophisticated methods can produce estimates of TF binding strength in a DNA segment, by integrating all putative binding sites, both weak and strong, present in that segment. Application of these methods to multiple *Drosophila* genomes, coupled with whole-genome alignments, can help describe the evolution of TF binding events. Cross species comparison can also improve the accuracy of predicting TF binding targets (6–8).

Computational methods have also been used to search for clusters of binding sites of multiple TFs, with the goal of identifying *cis*-regulatory modules (CRMs, also called enhancers). CRMs are ~500–1000-bp long regulatory elements that harbor multiple binding sites that together mediate a specific expression pattern of a neighboring gene (9). The identification of CRMs can provide a meaningful context in which the role of individual TF binding sites can be interpreted; they may also help reduce false positives in predicting individual binding sites. More recently, statistical methods have been demonstrated to recover functional CRMs without the prior knowledge of relevant TFs and/or their motifs. Such motif-blind approaches adopt the alternative paradigm of 'supervised CRM discovery', where a set of known CRMs with similar functionality (expression patterns) are used as 'training data' to locate other similar CRMs in the genome (10,11).

Genome Surveyor 2.0 presents an easy-to-use, web-based graphical interface to many of the *cis*-regulatory analysis tools mentioned above. It allows the user to perform TF target prediction and CRM discovery using any motif(s) from the FlyFactorSurvey database (12), the most comprehensive resource for *Drosophila* motifs today. It displays genome browser 'tracks' that profile matches to individual motifs or user-selected combinations of motifs, based on sequence information from a single genome or a

*To whom correspondence should be addressed. Tel: +1 217 333 3233; Fax: +1 217 265 6494; Email: sinhas@illinois.edu

combination of genomes. It also provides tracks for 'supervised CRM prediction' (10), driven by a user-selected subset of known CRMs from the REDfly database (13). Additional tracks are available to visualize related information such as chromatin immunoprecipitation (ChIP)-based profiles of TF occupancy, and previously characterized CRMs from the literature. In addition to providing locus-centric visualization of *cis*-regulatory elements, Genome Surveyor 2.0 provides an interface to search for motif/ChIP-based binding site clusters genome-wide.

WEB INTERFACE

Genome surveyor 2.0 provides users with the following components to perform *cis*-regulatory analysis in *Drosophila melanogaster* (Figure 1A).

(1) Single/multi-species motif profiles. A motif profile displays the estimated binding site presence for a user-selected TF motif as a function of genomic coordinates. We obtain the single species profiles by running the program Stubb (14) and multi-species profiles by averaging the profiles of orthologous regions from selected species.
(2) Supervised CRM discovery profiles. This component allows the user to specify a set of known CRMs and search for novel CRMs that have a similar *k-mer* composition to the specified set. Supervised CRM discovery methods do not require pre-selection of motifs, and provide a viable alternative to predicting functional CRMs, as explained in (10).
(3) Profiles of other cis-regulatory information. ChIP-based-binding profiles (from BDTNP) and experimentally validated CRMs (from REDfly) can be displayed along with other profiles. In addition to Stubb-based motif profiles, the user may visualize binding site predictions by a more traditional method (individual matches above a threshold).
(4) Search for Motif/ChIP clusters of binding sites. This component provides the user with an ability to search the entire genome (or list of loci) for the most significant clusters of motif matches and/or ChIP sites.

The first three components are implemented as plugins for GBrowse (15), and their outputs are 'tracks' that may be added to the current view of GBrowse. Note that all of these tracks/profiles can be displayed simultaneously, as illustrated in Figure 1B.

Single/multi-species motif profiles

We have pre-computed the motif profiles of a large collection of experimentally validated TFs for *D. melanogaster* (12) using the Hidden Markov Model-based program Stubb (14). (Stubb examines each 500-bp window and computes a score for the presence of one or more strong or weak binding sites in that window, without imposing arbitrary thresholds on what constitute a motif match.) We have also generated motif profiles for 11 other *Drosophila* species and mapped them to the *D.mel* coordinates. All profiles are normalized using their genome-wide mean and standard deviation. Users may select from the following options related to motif profiles:

- Individual species, individual motif: This option displays the profiles of the selected motif(s) in the selected species. Given this option, users might easily check, for example, whether a specific potential binding event is conserved between *D.mel* and *D.pse* by turning on the tracks of the corresponding motif for both species. Also, they may easily assess the similarity between the targets of two or more TFs. All tracks are directly linked to (and just a click away from) the FlyFactorSurvey database (12) that provides detailed information about the binding site's specificity and the method used to characterize it.
- Individual species, multi-motif: This option averages the profiles of selected motifs for each selected species. This provides a convenient way to look for clusters of binding sites of several TFs, as a means to discover novel CRMs. For example, a user searching for enhancers regulating dorsal/ventral (D/V) patterning may choose to select the motifs involved in this process (e.g. those for the TFs *Dl, Twi, Sna*) and examine their average profile. The user may repeat this process for other species as well, to examine if the predicted CRM in *D. melanogaster* is independently supported by predictions at orthologous locations in those species.
- Multi-species, individual motif: This option combines the profiles of a selected motif from different species, using simple averaging or a phylogenetic tree-based averaging (7). The peaks in this profile represent the TF targets that are conserved across species.
- Multi-species, multi-motif: This option averages all the profiles from selected motifs and species to create a single track. The peaks in this profile represent the strong clusters of binding sites that are conserved across species, and may thus correspond to functional CRMs.
- User-defined motif: This option allows users to input their own Position Weight Matrices (PWMs), rather than selecting from a pre-defined list of motifs. Although there has been an intense effort to characterize the binding specificities of all TFs in *D. melanogaster* (16), there remain many TFs with unknown binding specificity. The *user-defined motif* option allows motifs that are not part of the publically available database to be used.

Supervised CRM discovery profiles

The REDfly database catalogs over 800 experimentally characterized CRMs in *D. melanogaster*, along with their spatial/temporal expression patterns (13). This extensive resource can be used as 'training data' to computationally predict novel CRMs genome-wide, through 'supervised CRM discovery' methods. These methods score a genomic segment for sequence similarity to any given set

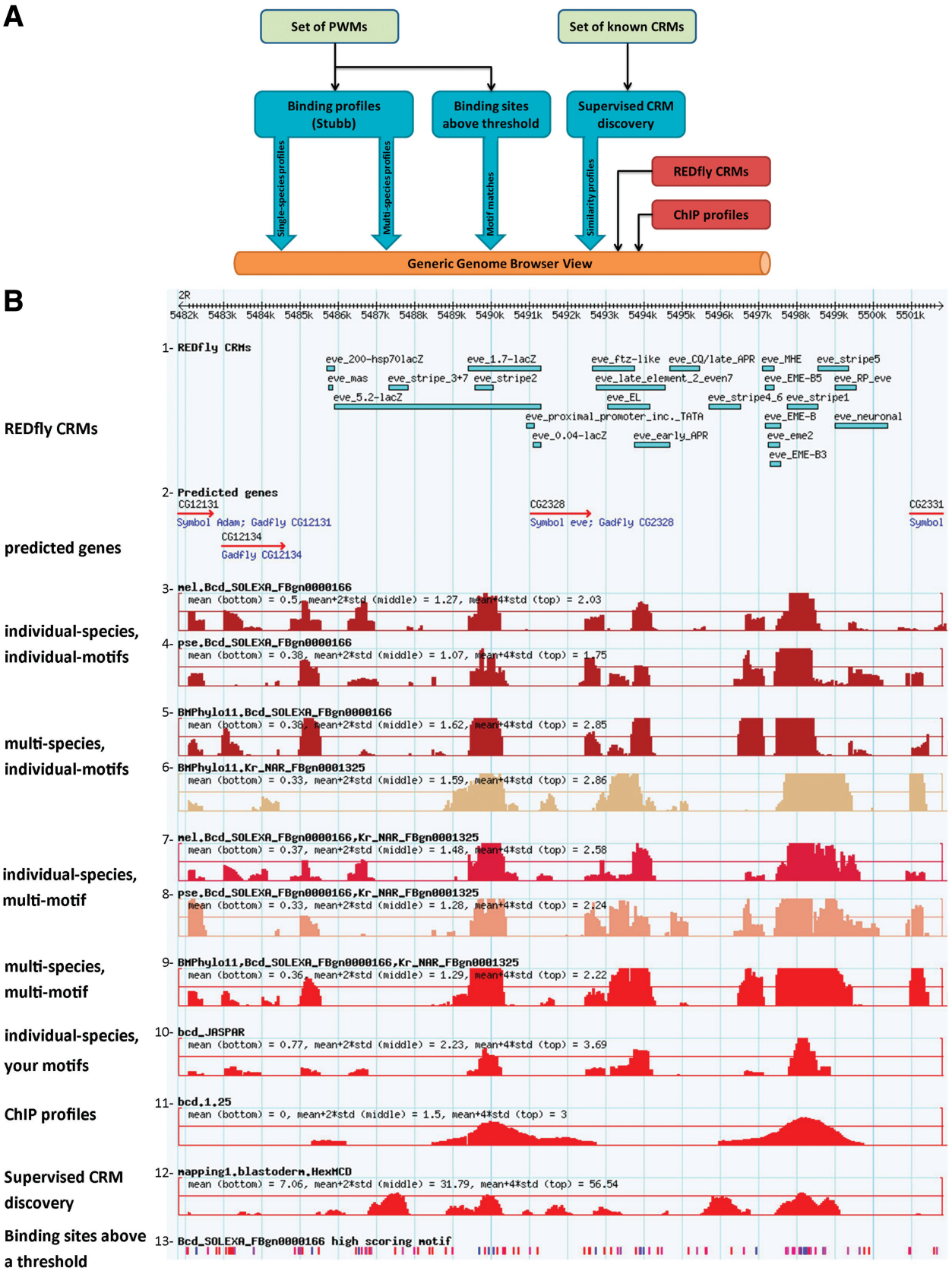
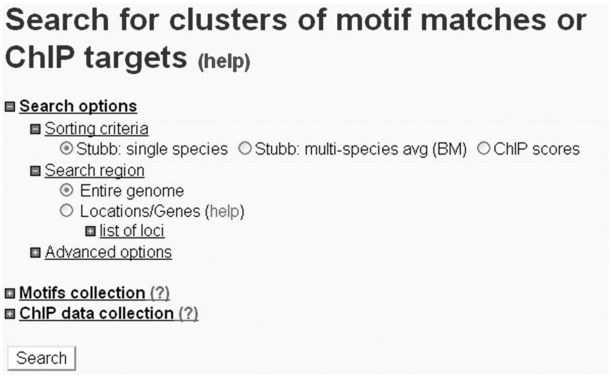


Figure 1. Genome Surveyor 2.0 input/output. (**A**) General Scheme of Genome Surveyor 2.0. Users may investigate the region of interest (defined through GBrowse) for potential *cis*-regulatory activity using a variety of tracks. To find the targets of specific motifs, users may activate the corresponding single species, multi-species and/or ChIP profiles. They may search for CRMs that are similar to a set of previously validated CRMs, using the 'Supervised CRM discovery' tracks. The location of strong binding sites in a region can be visualized with the 'Binding sites

(continued)

A



B

Index	Location	Stubb(SS)	**Stubb(MS)**	Separate scores	Neighborhood Info
1.	2R:5498200 eve_stripe1	5.18	8.74	**Motif Name** (SS) (MS): Kr_NAR_FBgn0001325 9.09 16.05; hb_SOLEXA_5_FBgn0001180 -0.88 -1.21; tll_NAR_FBgn0003720 4.86 7.62; hkb_NAR_FBgn0001204 5.69 8.1; Bcd_SOLEXA_FBgn0000166 6.27 11.94	1. Gene CG2331 TER94 (distance 2774, upstream) 2. Gene CG2328 eve (distance 5608, downstream)
2.	3L:8644850	3.32	7.35	**Motif Name** (SS) (MS): Kr_NAR_FBgn0001325 4.79 10.47; hb_SOLEXA_5_FBgn0001180 5.74 13.66; tll_NAR_FBgn0003720 2.67 5.39; hkb_NAR_FBgn0001204 2.88 5.63; Bcd_SOLEXA_FBgn0000166 0.5 1.63	1. Gene CG6494 h (distance 4856, upstream) 2. Gene CG6486 (distance 20187, downstream)

Figure 2. Search tool for motif/ChIP-based binding site clusters. (**A**) An interface to search for motif/ChIP-based binding site clusters. There are three main panels: *Search options, Motif collection* and *ChIP data collection*. Search options: here, the user may select the types of profiles ('Sorting criteria') to use in the search and also the search region, which is either the entire genome or a list of genomic loci. More advanced search options are also provided (e.g. the number of top results). Motifs collection and ChIP data collection: here, the user may select the specific motifs and ChIP profiles to use in the search. (**B**) Sample output of the search tool. A genome-wide search was conducted with five TFs involved in anterior patterning (BCD, KR, HB, TLL and HKB), retrieving segments (of 500 bp each) where a significant cluster of binding sites for one or more of these TFs was found. (Only the top two reported segments are shown in the figure.) The second column reports the location of the segment, along with a link to the Genome Surveyor view of this locus. It also provides the names of known CRMs that overlap the segment (e.g. 'eve_stripe1′ for segment 1). The third and fourth columns show the strength (*z*-score relative to the genome-wide average) of binding site clustering in the segment, based on single species ['Stubb(SS)'] and multi-species ['Stubb(MS)'] analysis respectively. The fifth column reports the scores of each selected motif, separately for single species ('SS') and multi-species ('MS') analysis. The last column provides information about the nearest neighboring genes of the segment.

of known CRMs. The similarity score is based on frequencies of short words in the sequences, and can detect the presence of shared binding sites without relying on prior knowledge of motifs. As such, this is a pragmatic approach to CRM discovery when the likely transcriptional regulators of a gene are not known in advance, or their binding specificities have not been characterized. Genome Surveyor 2.0 allows the user to profile any genomic region with two different scores [HexMCD and IMM (M. Kazemian, Q. Zhu, M. S. Halfon, S. Sinha, manuscript under preparation)] (10). The training set of CRMs may be selected as one of over 30 different subsets of REDfly CRMs, defined by the tissue/stage of development that they help regulate (11). The user may also upload a Fasta file of CRM sequences.

Figure 1. Continued
above threshold' track. Known CRMs are displayed through 'REDfly CRMs' track. (Green, red and blue boxes show inputs, built-in tracks and tracks that are computed on-the-fly, respectively.) (**B**) An instance of Genome Surveyor output. Shown is the 20 k region surrounding the *eve* gene. Thirteen different tracks are shown: (numbered from 1 to 13 top to bottom) [1]REDfly CRMs and [2]Predicted genes. [3,4]Motif profiles for *Bicoid* (Bcd_SOLEXA) in *D.mel* and *D.pse* as two separate tracks (selected from the 'Stubb: individual motifs, individual species' plugin). [5,6]Multi-species profiles for *Bicoid* (Bcd_SOLEXA) and *Kruppel* (Kr_NAR) motifs (selected from the 'Stubb: individual motifs, multi-species avg.' plugin). [7,8]Average of motif profiles for *Bicoid* (Bcd_SOLEXA) and *Kruppel* (Kr_NAR) motifs in *D.mel* and *D.pse* as two separated tracks (selected from the 'Stubb: multi motif avg., individual species' plugin). [9]Average of multi-species profiles for the Bcd_SOLEXA and Kr_NAR motifs (selected from the 'Stubb: multi-motif, multi-species avg.' plugin). [10]Motif profile for user provided PWM; here, the *Bicoid* (bcd) motif from the JASPAR database (selected from the 'Stubb: your motifs' plugin). [11]ChIP binding profile for *Bicoid* (selected from the 'ChIP Tracks' plugin). [12]The similarity profile to a set of known CRMs active in blastoderm-stage development (selected from the 'Supervised CRM discovery' plugin). [13]Individual high scoring sequence motif matches (blue codes for the highest score) for Bcd_SOLEXA (selected from the 'Binding sites above threshold').

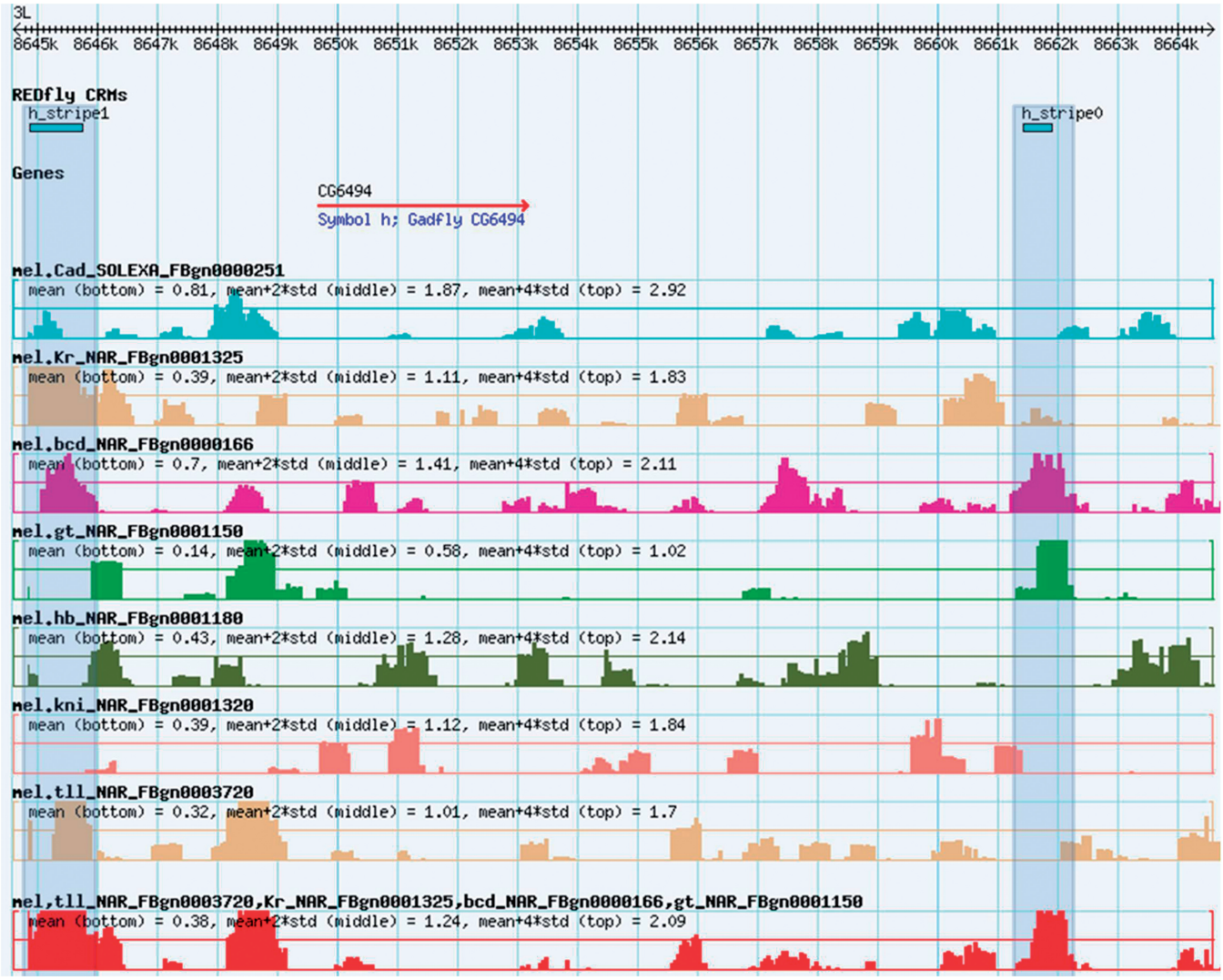


Figure 3. Example of motif composition analysis in known CRMs. Shown is the 21-kb region surrounding the *hairy* gene. Eight different tracks are shown: the 'Stubb individual species' profiles for seven TFs (CAD, KR, BCD, GT, HB, KNI, TLL) involved in initial stages of anterior–posterior (A/P) patterning in the embryo, and the 'Stubb multi-motif avg.' profile over TLL, KR, BCD and GT. Blue shaded regions represent two known A/P related CRMs, *h*_stripe1 and *h*_stripe0, that are both known to be BCD-activated CRMs (21). The presence (high peaks in score profile) of this known activator, as well as specific transcriptional repressors (TLL and KR for h_stripe1 and GT for h_stripe0), is consistent with the CRM expression patterns. The bottom profile (red) illustrates how combining Stubb profiles from multiple motifs might help in CRM discovery.

Binding sites, ChIP and REDfly profiles

Users may select from the following three tracks for additional information to aid their analysis:

- Binding sites above a threshold. This functionality is taken from (http://gmod.org/wiki/MotifFinder.pm). It displays individual binding sites predicted based on how well they match the selected/provided motifs.
- ChIP profiles. This track displays ChIP-based measurements of TF occupancy (17). At this time, these profiles are available for a limited number of TFs.
- REDfly CRMs. This track shows experimentally verified *D. melanogaster* CRMs from REDfly (13). It helps user to check the availability of any known enhancer in their region of interest. Each CRM is linked back to the REDfly database for detailed information (e.g. CRM expression pattern, the evidence for the element, the source, binding sites).

Search interface for Motif/ChIP clusters of binding sites

CRMs are known to harbor binding sites for several TFs, which act together to achieve specific regulatory functions. As such, computational tools for genome-wide CRM discovery typically search for clusters of binding sites with suitably chosen collections of TF motifs. Genome Surveyor 2.0 provides an interface for users to search for the most significant clusters of binding sites in the *D. melanogaster* genome for any user-specified combination of TFs (Figure 2A).

The search interface may be accessed from the main page of Genome Surveyor 2.0. Users first select the type of binding site profiles that will be used for search (Single/multi species motif profiles or ChIP profiles). Next, they may choose to scan the entire genome, or provide a list of genomic loci where the search will be performed. Advanced options (e.g. the number of top hits or the minimum number of different TFs in a predicted cluster) are available, but default settings are provided and help pages provide guidance for changing them. Finally, the user selects the motif or ChIP profiles of interest and begins the search. The output of the search tool is a table of predicted regulatory sequences (500-bp segments with clusters of binding sites) in the *D. melanogaster* genome, with links to appropriate GBrowse views (Figure 2B). The results are sorted based on the average value of the selected profiles in the segments. Single as well as multi-species scores are reported for each segment. Moreover, a score representing each motif's presence in the segment is shown separately, to help the user determine which motifs contribute significantly to the cluster. The output also includes information about the nearest neighboring genes and their distances from the binding site cluster.

Methods validation

Stubb is a popular CRM discovery tool that has been tested by multiple groups in different species (18–20). We have shown previously that regions with high Stubb scores are highly enriched for experimentally observed TF binding (ChIP), and that the enrichment improves significantly upon incorporating multi-species information (7). Stubb score profiles can be utilized to investigate the binding site composition of any genomic region. Figure 3 shows an example of motif regulatory analysis for two known CRMs. The strategy of combining the Stubb profiles of multiple TFs and identify the segments with highest average scores (Figure 3) has been demonstrated to recover known CRMs (16). Genome-wide predictions of the 'supervised CRM prediction' methods included in Genome Surveyor 2.0 have been assessed statistically and validated experimentally (10).

FUNDING

Funding for open access charge: This work was supported in part by grants by the National Institute of Health (grant R01HG004744-01 to M.H.B., grant R01GM085233-01 to S.S.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W., Iyer,V.N. *et al*. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
2. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.
3. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al*. (2009) FlyBase: enhancing Drosophila gene ontology annotations. *Nucleic Acids Res*., **37**, D555–D559.
4. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al*. (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
5. The modENCODE Consortium, Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L. *et al*. (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.
6. Berman,B.P., Pfeiffer,B.D., Laverty,T.R., Salzberg,S.L., Rubin,G.M., Eisen,M.B. and Celniker,S.E. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol*., **5**, R61.
7. Kazemian,M., Blatti,C., Richards,A., McCutchan,M., Wakabayashi-Ito,N., Hammonds,A.S., Celniker,S.E., Kumar,S., Wolfe,S.A., Brodsky,M.H. *et al*. (2010) Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. *PLoS Biol*., **8**, e1000456.
8. Sinha,S., Schroeder,M.D., Unnerstall,U., Gaul,U. and Siggia,E.D. (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, **5**, 129.
9. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*, 1st edn. Academic Press, Burlington, MA.
10. Kantorovitz,M.R., Kazemian,M., Kinston,S., Miranda-Saavedra,D., Zhu,Q., Robinson,G.E., Gottgens,B., Halfon,M.S. and Sinha,S. (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Dev. Cell*, **17**, 568–579.
11. Ivan,A., Halfon,M.S. and Sinha,S. (2008) Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol*., **9**, R22.
12. Zhu,L.J., Christensen,R.G., Kazemian,M., Hull,C.J., Enuameh,M.S., Basciotta,M.D., Brasefield,J.A., Zhu,C., Asriyan,Y., Lapointe,D.S. *et al*. (2011) FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res*, **39**, D111–D117.
13. Gallo,S.M., Gerrard,D.T., Miner,D., Simich,M., Des Soye,B., Bergman,C.M. and Halfon,M.S. (2010) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res*., **39**, D118–D123.
14. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19(Suppl. 1)**, i292–i301.
15. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
16. Noyes,M.B., Meng,X., Wakabayashi,A., Sinha,S., Brodsky,M.H. and Wolfe,S.A. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res*., **36**, 2547–2560.
17. MacArthur,S., Li,X.Y., Li,J., Brown,J.B., Chu,H.C., Zeng,L., Grondona,B.P., Hechmer,A., Simirenko,L., Keranen,S.V. *et al*. (2009) Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol*., **10**, R80.
18. Won,K.J., Agarwal,S., Shen,L., Shoemaker,R., Ren,B. and Wang,W. (2009) An integrated approach to identifying cis-regulatory modules in the human genome. *PLoS ONE*, **4**, e5501.

19. Su,J., Teichmann,S.A. and Down,T.A. (2010) Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol.*, **6**, e1001020.
20. Siddharthan,R. (2008) PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Comput. Biol.*, **4**, e1000156.
21. Ochoa-Espinosa,A., Yucel,G., Kaplan,L., Pare,A., Pura,N., Oberstein,A., Papatsenko,D. and Small,S. (2005) The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc. Natl Acad. Sci. USA*, **102**, 4960–4965.